

# DISTRIBUTION OF THE LARGEST ROOT OF A MATRIX FOR ROY'S TEST IN MULTIVARIATE ANALYSIS OF VARIANCE

BY MARCO CHIANI

*University of Bologna*

Let  $\mathbf{X}, \mathbf{Y}$  denote two independent real Gaussian  $p \times m$  and  $p \times n$  matrices with  $m, n \geq p$ , each constituted by zero mean independent, identically distributed columns with common covariance. The Roy's largest root criterion, used in multivariate analysis of variance (MANOVA), is based on the statistic of the largest eigenvalue,  $\Theta_1$ , of  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$ , where  $\mathbf{A} = \mathbf{X}\mathbf{X}^T$  and  $\mathbf{B} = \mathbf{Y}\mathbf{Y}^T$  are independent central Wishart matrices. We derive a new expression and efficient recursive formulas for the exact distribution of  $\Theta_1$ . The expression can be easily calculated even for large parameters, eliminating the need of pre-calculated tables for the application of the Roy's test.

**1. Introduction.** The joint distribution of  $s$  non-null eigenvalues of a multivariate real beta matrix in the null case can be written in the form [14, page 112], [2, page 331],

$$(1.1) \quad f(\theta_1, \dots, \theta_s) = C(s, m, n) \prod_{i=1}^s \theta_i^m (1 - \theta_i)^n \cdot \prod_{i < j}^s (\theta_i - \theta_j)$$

where  $1 > \theta_1 \geq \theta_2 \geq \dots \geq \theta_s > 0$ , and  $C(s, m, n)$  is a normalizing constant given by

$$(1.2) \quad C = C(s, m, n) = \pi^{s/2} \prod_{i=1}^s \frac{\Gamma\left(\frac{i+2m+2n+s+2}{2}\right)}{\Gamma\left(\frac{i}{2}\right) \Gamma\left(\frac{i+2m+1}{2}\right) \Gamma\left(\frac{i+2n+1}{2}\right)}.$$

This distribution arises in multivariate analysis of variance (MANOVA) and, with the notation introduced above, is the distribution of the eigenvalues of  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$  with parameters

$$(1.3) \quad s = p, \quad m = (n - p - 1)/2, \quad n = (m - p - 1)/2.$$

The marginal distribution of the largest eigenvalue,  $\Theta_1$ , is of basic importance in testing hypotheses and constructing confidence regions in multivariate analysis of

---

AMS 2000 subject classifications: Primary 62H10; secondary 15A52

Keywords and phrases: Roy's test, Random Matrices, multivariate analysis of variance (MANOVA), characteristic roots, largest eigenvalue, Tracy-Widom distribution, Wishart matrices.

variance (MANOVA) according to the Roy's largest root criterion (see e.g. [2, page 333] and references therein), and is generally considered difficult to compute. For this reason, extensive studies have produced tables of upper percentage points for few specific (small) values of  $s$  and some combinations of  $m, n$  (see e.g. [8, 15], [2, Table B.4]). The most efficient numerical algorithm to compute the cumulative distribution function (CDF) of  $\Theta_1$  is provided in [3] based on [7], but for nonintegers  $m$  or  $n$  it requires infinite series expansions which can results in computational time of several hours. Approximations based on the Tracy-Widom distribution are discussed in [12].

In this paper we derive a simple expression for the exact CDF of  $\Theta_1$  for arbitrary  $s, n, m$ , and an iterative algorithm for its fast evaluation. The algorithm needs only the incomplete beta function, and does not rely on numerical integration or series expansion. For instance, all results in [2, Table B.4] or in [12, Table 1] can be easily computed instantaneously<sup>1</sup>; even for the most challenging cases analyzed in literature [3, Table 1], which with previous methods required hours of computational time, no more than a second is needed. Finally, we discuss some approximations based on the Tracy-Widom distribution and its approximation [12, 4].

We remark that we study the case of real matrices: the complex analogous of our problem, i.e., the case where  $\mathbf{X}, \mathbf{Y}$  are independent complex Gaussian, is much easier and has been solved in [13].

Throughout the paper we indicate with  $\Gamma(\cdot)$  the gamma function, with  $B(a, b)$  the beta function, with  $\mathcal{B}(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$  the incomplete (lower) beta function [1, Ch. 6], and with  $|\cdot|$  the determinant.

## 2. Exact distribution of the largest eigenvalue for multivariate beta matrices in the null case.

The following is the main result of the paper.

**THEOREM 1.** *The CDF of the largest eigenvalue  $\Theta_1$  for a multivariate beta matrix in the null case is:*

$$(2.1) \quad F_{\Theta_1}(\theta_1) = \Pr \{ \Theta_1 \leq \theta_1 \} = C \sqrt{|\mathbf{A}(\theta_1)|}.$$

When  $s$  is even, the elements of the  $s \times s$  skew-symmetric matrix  $\mathbf{A}(\theta_1)$  are:

$$(2.2) \quad a_{i,j}(\theta_1) = \mathcal{E}(\theta_1; m+j, m+i) - \mathcal{E}(\theta_1; m+i, m+j) \quad i, j = 1, \dots, s$$

where

$$(2.3) \quad \mathcal{E}(x; a, b) \triangleq \int_0^x t^{a-1}(1-t)^n \mathcal{B}(t; b, n+1) dt.$$

<sup>1</sup>On a current desktop computer in less than 0.1 seconds.

When  $s$  is odd, the elements of the  $(s+1) \times (s+1)$  skew-symmetric matrix  $\mathbf{A}(x_1)$  are as in (2.2), with the additional elements

$$(2.4) \quad a_{i,s+1}(\theta_1) = \mathcal{B}(\theta_1; m+i, n+1) \quad i = 1, \dots, s$$

$$(2.5) \quad a_{s+1,j}(\theta_1) = -a_{j,s+1}(\theta_1) \quad j = 1, \dots, s$$

$$(2.6) \quad a_{s+1,s+1}(\theta_1) = 0$$

Note that  $a_{i,j}(\theta_1) = -a_{j,i}(\theta_1)$  and  $a_{i,i}(\theta_1) = 0$ .

Moreover, the elements  $a_{i,j}(\theta_1)$  can be computed iteratively, starting from the beta function, without numerical integration or series expansion.

PROOF. The proof is based on the approach introduced in [4] for Wishart and GOE matrices.

Denoting  $\xi(x) = x^m(1-x)^n$ ,  $\mathbf{x} = [x_1, x_2, \dots, x_s]$ , and with  $\mathbf{V}(\mathbf{x}) = \{x_j^{i-1}\}$  the Vandermonde matrix, we have for the eigenvalues in ascending order

$$(2.7) \quad f(x_s, \dots, x_1) = C \prod_{i < j} (x_j - x_i) \prod_{i=1}^s \xi(x_i) = C |\mathbf{V}(\mathbf{x})| \prod_{i=1}^s \xi(x_i)$$

where now  $0 < x_1 \leq \dots \leq x_s < 1$ .

The CDF of the largest eigenvalue is then

$$(2.8) \quad F_{\Theta_1}(\theta_1) = \int \dots \int_{0 \leq x_1 < \dots < x_s \leq \theta_1} f(x_s, \dots, x_1) d\mathbf{x}$$

$$(2.9) \quad = C \int \dots \int_{0 \leq x_1 < \dots < x_s \leq \theta_1} |\mathbf{V}(\mathbf{x})| \prod_{i=1}^s \xi(x_i) d\mathbf{x}.$$

To evaluate this integral we recall that for a generic  $s \times s$  matrix  $\Phi(\mathbf{x})$  with elements  $\{\Phi_i(x_j)\}$  the following identity holds [6]

$$(2.10) \quad \int \dots \int_{a \leq x_1 < \dots < x_s \leq b} |\Phi(\mathbf{x})| d\mathbf{x} = \text{Pf}(\mathbf{A})$$

where  $\text{Pf}(\mathbf{A}) = \sqrt{|\mathbf{A}|}$  is the Pfaffian, and the skew-symmetric matrix  $\mathbf{A}$  is  $s \times s$  for  $s$  even, and  $(s+1) \times (s+1)$  for  $s$  odd, with

$$(2.11) \quad a_{i,j} = \int_a^b \int_a^b \text{sgn}(y-x) \Phi_i(x) \Phi_j(y) dx dy \quad i, j = 1, \dots, s.$$

For  $s$  odd the additional elements are  $a_{i,s+1} = -a_{s+1,i} = \int_a^b \Phi_i(x)dx$ ,  $i = 1, \dots, s$ , and  $a_{s+1,s+1} = 0$ .

Using (2.10) in (2.9) with  $a = 0, b = \theta_1, \Phi_i(x) = x^{i-1}\xi(x) = x^{i-1}x^m(1-x)^n$  with some simple manipulations gives Theorem 1.

To avoid the numerical integration in (2.3), we first observe that the incomplete beta functions can be computed iteratively by the relation

$$(2.12) \quad \mathcal{B}(x; a+1, b) = \frac{a}{a+b} \mathcal{B}(x; a, b) - \frac{x^a(1-x)^b}{a+b}.$$

This relation is obtained for example with integration by parts of  $\int t^a(1-t)^{b-1}dt = \int f(t)g'(t)dt$  with  $f(t) = t^a - t^{a+1}$ ,  $g'(t) = (1-t)^{b-2}$ .

Then, from (2.3) and (2.12) the following identities can be easily verified:

$$(2.13) \quad \mathcal{E}(x; a, a) = \frac{1}{2} \mathcal{B}(x; a, n+1)^2$$

$$(2.14) \quad \mathcal{E}(x; a, b+1) = b \frac{\mathcal{E}(x; a, b)}{b+n+1} - \frac{\mathcal{B}(x; a+b, 2n+2)}{b+n+1}$$

$$(2.15) \quad \mathcal{E}(x; b, a) = \mathcal{B}(x; a, n+1) \mathcal{B}(x; b, n+1) - \mathcal{E}(x; a, b).$$

Therefore, the elements of the matrix in (2.1) can be evaluated iteratively without any numerical integration.  $\square$

Theorem 1 can thus be used for an efficient computation of the exact CDF of the largest eigenvalue for Roy's test. In fact, using (2.13), (2.14), and (2.15), the CDF in (2.1) can be simply evaluated, without numerical integrations or series expansion, by Algorithm 1 reported below.

Implementing directly the algorithm in Mathematica on a personal computer, for each value  $x_1$  we obtain the exact CDF in (2.1) in less than 0.1 seconds for all tables in [15], [2, Table B.4] and [12, Table 1]. For  $s = 54, m = -1/2, n = 45/2$ , which corresponds to the most challenging case analyzed in literature [3, Table 1, last row], the exact CDF is computed with the new expression in less than a second. For comparison, the approach based on series expansions of the hypergeometric function of a matrix argument requires hours. For larger parameters for which no other methods are available in literature, like for example  $s = 200, m = -1/2, n = 299/2$ , the computational time is less than fifteen seconds on a common personal computer. Therefore, the upper percentage points for Roy's test can be evaluated exactly and almost instantaneously for parameters of interest in applied statistics.

**Algorithm 1** Algorithm: CDF of the largest eigenvalue for Roy's test**Input:**  $s, m, n, \theta_1$ **Needed:**Incomplete beta function  $\mathcal{B}(x; a, b)$ **Output:**  $F_{\Theta_1}(\theta_1) = \Pr\{\Theta_1 \leq \theta_1\}$  $\mathbf{A} = \mathbf{0}$ **for**  $i = 1 \rightarrow s$  **do** $b_i = \mathcal{B}(\theta_1; m + i, n + 1)^2 / 2$ **for**  $j = i \rightarrow s - 1$  **do** $b_{j+1} = \frac{m + j}{m + j + n + 1} b_j - \frac{\mathcal{B}(\theta_1; 2m + i + j, 2n + 2)}{m + j + n + 1}$  $a_{i,j+1} = \mathcal{B}(\theta_1; m + i, n + 1) \mathcal{B}(\theta_1; m + j + 1, n + 1) - 2b_{j+1}$ **end for****end for****if**  $s$  is odd **then**Append one column to  $\mathbf{A}$  with  $a_{i,s+1} = \mathcal{B}(\theta_1; m + i, n + 1)$ ,  $i = 1, \dots, s$ Append one zero row to  $\mathbf{A}$ **end if** $\mathbf{A} = \mathbf{A} - \mathbf{A}^T$ **return**  $F_{\Theta_1}(\theta_1) = C(s, m, n) \sqrt{|\mathbf{A}|}$ 

**3. Approximations.** In [12] it is shown that the logit of  $\Theta_1$  approaches the Tracy-Widom law for large  $s$ . More precisely, it is shown that, for  $m \geq -1/2$ ,  $n \geq 0$  and large  $s$ , the following approximation holds

$$(3.1) \quad \frac{\log(\Theta_1/(1 - \Theta_1)) - \mu}{\sigma} \stackrel{\mathcal{D}}{\approx} \mathcal{TW}_1$$

where  $\mathcal{TW}_1$  denotes a random variable with Tracy-Widom distribution of order 1 [16, 17, 9, 10, 18], and the values of  $\mu, \sigma$  are given, using the parameters in (1.3), by [11, 12]

$$(3.2) \quad \mu = 2 \log \tan \left( \frac{\gamma + \phi}{2} \right)$$

$$(3.3) \quad \sigma^3 = \frac{16}{(m + n - 1)^2 \sin^2(\gamma + \phi) \sin \gamma \sin \phi}$$

$$(3.4) \quad \gamma = \arccos \left( \frac{m + n - 2p}{m + n - 1} \right)$$

$$(3.5) \quad \phi = \arccos \left( \frac{m - n}{m + n - 1} \right).$$

Moreover, in [4] it is shown that the  $\mathcal{TW}_1$  can be closely approximated by a shifted gamma distribution, so that

$$(3.6) \quad F_1(x) \triangleq \Pr(\mathcal{TW}_1 \leq x) \simeq P\left(k, \frac{x + \alpha}{\delta}\right) \quad x > -\alpha$$

$$(3.7) \quad x = F_1^{-1}(y) \simeq \delta P^{-1}(k, y) - \alpha$$

where  $P(a, x)$  is the regularized lower incomplete gamma function,  $P^{-1}(a, y)$  is its inverse, and the constants  $k = 46.446, \delta = 0.186054, \alpha = 9.84801$  have been chosen to match the moments of the approximation to that of the Tracy-Widom. Thus, using (3.6) in (3.1) we obtain for the CDF of the Roy's statistic in the null case

$$(3.8) \quad F_{\Theta_1}(\theta_1) \simeq P\left(k, \frac{\log(\theta_1/(1 - \theta_1)) - \mu + \sigma\alpha}{\delta}\right)$$

and for its inverse, useful for evaluating the percentiles,

$$(3.9) \quad \theta_1 = F_{\Theta_1}^{-1}(y) \simeq \frac{\exp\{\sigma[\delta P^{-1}(k, y) - \alpha] + \mu\}}{1 + \exp\{\sigma[\delta P^{-1}(k, y) - \alpha] + \mu\}}.$$

An example comparing exact (by Algorithm 1) with approximate (by (3.8)) distributions is reported in Fig. 1.

While Theorem 1 with Algorithm 1 gives the exact distribution, the two approximations above, which are in simple closed forms, can be used for a rapid, approximate design of the Roy's test. For instance, with  $s = 5, m = -1/2, n = 1000$  the 80th-percentile obtained by root-finding with the exact CDF of Algorithm 1 is  $\theta_1 = 0.008501$ , while (3.9) gives  $\theta_1 = 0.008609$ . For larger matrices and larger values of the CDF the approximation generally improves. For example, with  $s = 200, m = -1/2, n = 299/2$  the 99th-percentile obtained by Algorithm 1 is  $\theta_1 = 0.827760$ , while (3.9) gives  $\theta_1 = 0.827761$ .

**4. Remarks for the complex multivariate beta.** For completeness we recall that, when  $\mathbf{X}, \mathbf{Y}$  are two independent *complex* Gaussian, the analogous of (1.1) is the complex multivariate beta, where the joint distribution of the eigenvalues is [13]

$$(4.1) \quad f(\theta_1, \dots, \theta_s) = C'(s, m, n) \prod_{i=1}^s \theta_i^m (1 - \theta_i)^n \cdot \prod_{i < j}^s (\theta_i - \theta_j)^2$$

where  $1 > \theta_1 \geq \theta_2 \geq \dots \geq \theta_s > 0$ , and

$$C'(s, m, n) = \prod_{i=1}^s \frac{\Gamma(m + n + s + i)}{\Gamma(i) \Gamma(i + m) \Gamma(i + n)}.$$

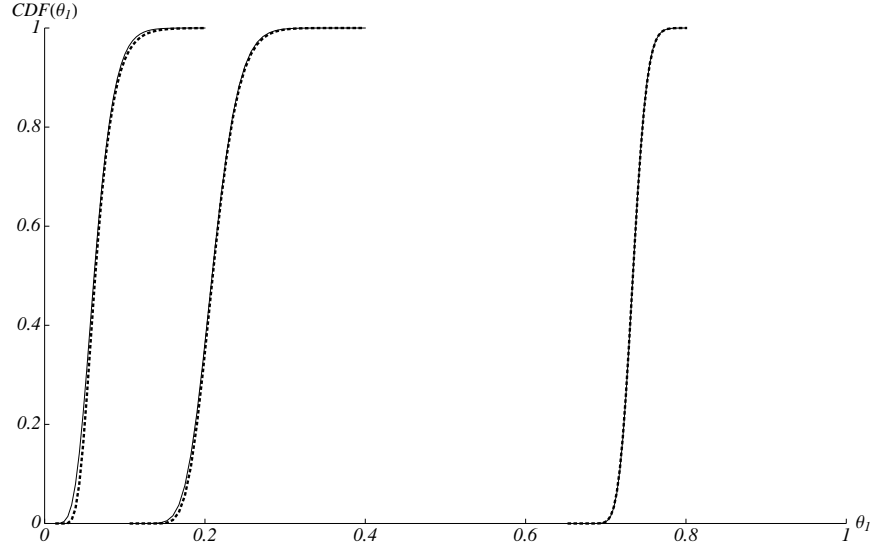


FIG 1. CDF of the largest eigenvalue, real beta matrix,  $m = -1/2, n = 100$ . From left to right:  $s = 5, 15, 100$ . Comparison between the exact distribution (2.1) (solid line) and the approximation in (3.8) (dotted). Note that for  $s = 100$  the two curves are almost indistinguishable.

In this case we can write

$$(4.2) \quad f(\theta_1, \dots, \theta_s) = C'(s, m, n) |\mathbf{V}(\theta)|^2 \prod_{i=1}^s \xi(\theta_i).$$

Therefore, by applying [5, Corollary 2] the CDF of  $\Theta_1$  in the complex case is simply given by [13]

$$(4.3) \quad F_{\Theta_1}(\theta_1) = \Pr \{ \Theta_1 \leq \theta_1 \} = C'(s, m, n) |\mathbf{A}(\theta_1)|$$

where the elements of the  $s \times s$  matrix  $\mathbf{A}(\theta_1)$  are:

$$(4.4) \quad a_{i,j}(\theta_1) = \mathcal{B}(\theta_1, m + i + j - 1, n + 1) \quad i, j = 1, \dots, s.$$

Moreover, expressions analogous to (3.6) and (3.9) can be derived, by using the Tracy-Widom limiting behavior for  $\Theta_1$  in the complex case [11] and the approximation of the Tracy-Widom of order 2 with a shifted gamma [4].

## REFERENCES

- [1] ABRAMOWITZ, M. and STEGUN, I. A. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. United States Department of Commerce, Washington, D.C.

- [2] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.
- [3] BUTLER, R. W. and PAIGE, R. L. (2011). Exact distributional computations for Roy's statistic and the largest eigenvalue of a Wishart distribution. *Statistics and Computing* **21** 147–157.
- [4] CHIANI, M. (2012). Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy-Widom distribution. *arXiv preprint arXiv:1209.3394*.
- [5] CHIANI, M., WIN, M. Z. and ZANELLA, A. (2003). On the Capacity of Spatially Correlated MIMO Rayleigh Fading Channels. *IEEE Trans. Inform. Theory* **49** 2363–2371.
- [6] DE BRUIJN, N. G. (1955). On some multiple integrals involving determinants. *J. Indian Math. Soc* **19** 133–151.
- [7] GUPTA, R. D. and RICHARDS, D. S. P. (1985). Hypergeometric functions of scalar matrix argument are expressible in terms of classical hypergeometric functions. *SIAM J. Math. Anal.* **16** 852–858.
- [8] HECK, D. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root. *The Annals of Mathematical Statistics* **31** 625–642.
- [9] JOHANSSON, K. (2000). Shape Fluctuations and Random Matrices. *Communications in Mathematical Physics* **209** 437–476.
- [10] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics* **29** 295–327.
- [11] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence. *The Annals of statistics* **36** 2638–2716.
- [12] JOHNSTONE, I. M. (2009). Approximate null distribution of the largest root in multivariate analysis. *The annals of applied statistics* **3** 1616–1633.
- [13] KHATRI, C. G. (1964). Distribution of the largest or the smallest characteristic root under null hypothesis concerning complex multivariate normal populations. *Ann. Math. Stat.* **35** 1807–1810.
- [14] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.
- [15] PILLAI, K. S. (1967). Upper percentage points of the largest root of a matrix in multivariate analysis. *Biometrika* **54** 189–194.
- [16] TRACY, C. A. and WIDOM, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics* **159** 151–174.
- [17] TRACY, C. A. and WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177** 727–754.
- [18] TRACY, C. A. and WIDOM, H. (2009). The distributions of random matrix theory and their applications. *New Trends in Mathematical Physics* 753–765.

DEI, UNIVERSITY OF BOLOGNA  
V.LE RISORGIMENTO 2, 40136  
BOLOGNA  
[marco.chiani@unibo.it](mailto:marco.chiani@unibo.it)